

A probabilistic proof of cutoff in the Metropolis algorithm for the Erdős-Rényi random graph

Winfried Barta

*University of Chicago
Department of Statistics*

December 20, 2011

Abstract

We consider the Metropolis algorithm for the distribution $\pi(x) = \theta^{S(x)}(1 + \theta)^{-n}$ on the hypercube $\mathcal{X} = \{0, 1\}^n$, where $S(x)$ is the number of ones in $x \in \{0, 1\}^n$ and $\theta \in (0, 1]$ is a constant. For $n = \binom{\nu}{2}$ this distribution corresponds to the Erdős-Rényi random graph model on ν vertices, where each edge is present independently with probability $\frac{\theta}{1+\theta}$. The lazy random walk Metropolis algorithm for this model specifies a Markov chain (X_t) on \mathcal{X} that is known to have cutoff at $\frac{1}{1+\theta}n \log n$ with window size n , a result derived by Fourier analysis in Diaconis and Hanlon (1992) and Ross and Xu (1994). In this work we give a new proof of this result that is purely probabilistic. This is done in the hope that probabilistic techniques will be easier to generalize to other, less symmetric distributions π . Our proof uses coupling and a projection to a two-dimensional Markov chain $X_t \rightarrow (S(X_t), d(X_0, X_t))$, where $d(X_0, \cdot)$ is the Hamming distance to the starting state X_0 .

1 Introduction

We are interested in analyzing convergence rates of the random walk Metropolis Hastings algorithm for various distributions π on the hypercube $\mathcal{X} := \{0, 1\}^n$. This Markov chain on \mathcal{X} moves as follows: Given that we are at a state $x \in \mathcal{X}$, we chose one of the n neighbors of x uniformly at random, say y , and propose to move from x to y . Here x is called a neighbor of y , denoted by $x \sim y$, whenever x and y differ in exactly one coordinate. This proposal gets accepted, i.e. we move to y , with probability $\min(1, \frac{\pi(y)}{\pi(x)})$. If it gets rejected we stay at x . This transition rule ensures that we have detailed balance

$$\pi(x) p(x, y) = \pi(y) p(y, x) \quad \text{for all } x, y \in \mathcal{X},$$

and since the chain is irreducible, its unique stationary distribution is π .

Often it is convenient to make the chain *lazy*. This corresponds to flipping a fair coin independently at each step. If it comes up heads, we stay where we are; if it comes up tails, we move according to the rule specified above. If \mathbb{P} is the transition probability matrix of the original chain, then the lazy chain has transition probability matrix $\mathbb{P}' := \frac{\mathbb{P}+I}{2}$, where I is the identity matrix.

All the stationary distributions π we will investigate are going to be *strictly unimodal*. By this we mean that there exists a state, z say, that has higher mass under π than any other state, and the π -mass decreases strictly with distance to z , i.e.

$$\pi(x) < \pi(y) \quad \text{whenever } d(x, z) > d(y, z).$$

Here, $d(x, z) := \sum_{i=1}^n |x_i - z_i|$ is the graph distance of x and z , i.e the number of coordinates where x and z differ. We also require π to be *radially symmetric* (with respect to the mode z), i.e.

$$\pi(x) = \pi(y) \quad \text{whenever } d(x, z) = d(y, z).$$

By relabeling the states we may (and will) assume that the mode z of π is at $\underline{0} = (0, 0, \dots, 0)$. This ensures that π is constant on level sets $L(k) := \{x \in \mathcal{X} : S(x) = k\}$ where $k \in \{0, 1, \dots, n\}$ and $S(x) := \sum_{i=1}^n x_i$ is the number of ones in x . Hence, in the Metropolis algorithm we will always accept downward moves $x \rightarrow y$ where $S(x) > S(y)$, and we will accept upward moves $x \rightarrow y$ where $S(x) < S(y)$ with probability $\theta_{S(x)}$, where we write $\theta_k := \frac{\pi(v)}{\pi(w)}$ where v, w are some (any) states such that $S(v) = k + 1$ and $S(w) = k$.

For concreteness, the transition kernel for the lazy random walk Metropolis Hastings algorithm is

$$p(x, y) = \begin{cases} \frac{1}{2n} & : x \sim y, S(x) > S(y), \\ \frac{1}{2n} \theta_{S(x)} & : x \sim y, S(x) < S(y), \\ \frac{1}{2} + \frac{n-S(x)}{2n} (1 - \theta_{S(x)}) & : x = y, \\ 0 & : \text{otherwise.} \end{cases} \quad (1)$$

Note that a consequence of radial symmetry is that the projection $S(X_t)$ of the Markov chain (X_t) is also Markov, since for all x, y we then have

$$P(x, [y]) = P(x', [y]) \quad \text{for all } x' \sim_S x.$$

Here we write $x \sim_S y$ whenever $S(x) = S(y)$ and $[y] := \{z : S(z) = S(y)\}$ for $y \in \mathcal{X}$ are the equivalence classes of the relation \sim_S .

We measure distance to stationarity by total variation:

$$\|P^t(x, \cdot) - \pi\|_{TV} := \max_{A \subset \mathcal{X}} (P^t(x, A) - \pi(A)),$$

and we are interested in this distance from the worst starting point:

$$d(t) := \max_{x \in \mathcal{X}} \|P^t(x, \cdot) - \pi\|_{TV}.$$

The mixing time for a parameter $\varepsilon \in (0, 1)$ is defined as

$$t_{mix}(\varepsilon) := \min\{t \geq 0 : d(t) \leq \varepsilon\}$$

and we write t_{mix} for $t_{mix}(1/4)$. We are interested in the behavior of t_{mix} as n goes to infinity.

An interesting phenomenon is that for some chains the mixing time $t_{mix}(\varepsilon)$ doesn't depend on the parameter ε (asymptotically, as n goes to infinity). We say a sequence $(\mathbf{X}^{(n)})_{n \in \mathbb{N}}$ of Markov chains $\mathbf{X}^{(n)} = (X_t^{(n)})_{t=0,1,\dots}$ on $\{0,1\}^n$ has a *cutoff* (at $t_{mix}^{(n)}$) if, for all $\varepsilon > 0$,

$$\lim_{n \rightarrow \infty} \frac{t_{mix}^{(n)}(\varepsilon)}{t_{mix}^{(n)}(1-\varepsilon)} = 1. \quad (2)$$

Here, $t_{mix}^{(n)}(\varepsilon)$ denotes the ε -mixing time of the n^{th} chain $(X_t^{(n)})_{t=0,1,\dots}$. This is equivalent to

$$\lim_{n \rightarrow \infty} d_n(c t_{mix}^{(n)}) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1. \end{cases} \quad (3)$$

So the function $d_n(\cdot)$, the total variation distance to stationarity from the worst starting point for the n^{th} chain, approaches a step function as n goes to infinity (if we rescale time by $t_{mix}^{(n)}$). For a proof of this equivalence see Levin et al. (2009, Lemma 18.1, page 247), from where we also borrowed the notation. For an overview of this cutoff phenomenon, see Diaconis (1996).

The following result is well known and not hard to prove:

Proposition 1.1. *For a sequence of finite Markov chains, the following are equivalent:*

- (i) *The sequence has a cutoff at $t_{mix}^{(n)}$.*
- (ii) *For all $\varepsilon \in (0, 1)$ we have $\lim_{n \rightarrow \infty} \frac{t_{mix}^{(n)}(\varepsilon)}{t_{mix}^{(n)}(1-\varepsilon)} = 1$.*
- (iii) *For all $\varepsilon \in (0, 1)$ we have $t_{mix}^{(n)}(\varepsilon) = t_{mix}^{(n)} \cdot [1 + h(n, \varepsilon)]$ for some $h(n, \varepsilon)$ with $\lim_{n \rightarrow \infty} h(n, \varepsilon) = 0$.*
- (iv) *For all $\varepsilon, \varepsilon' \in (0, 1)$ we have $t_{mix}^{(n)}(\varepsilon) \sim t_{mix}^{(n)}(\varepsilon')$.*
- (v) $\lim_{n \rightarrow \infty} d_n(c t_{mix}^{(n)}) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1. \end{cases}$
- (vi) $\lim_{n \rightarrow \infty} d_n(c t_{mix}^{(n)}(\varepsilon)) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1, \end{cases}$ for some $\varepsilon \in (0, 1)$.

$$(vii) \lim_{n \rightarrow \infty} d_n(c t_{mix}^{(n)}(\varepsilon)) = \begin{cases} 1 & \text{if } c < 1, \\ 0 & \text{if } c > 1, \end{cases} \quad \text{for all } \varepsilon \in (0, 1).$$

Sometimes we can analyze more precisely what happens for $c = 1$ in (3). We say a sequence of Markov chains has a cutoff with *window size* (w_n) , if $w_n \in o(t_{mix}^{(n)})$ and

$$\lim_{\alpha \rightarrow \infty} \liminf_{n \rightarrow \infty} d_n(t_{mix}^{(n)} - \alpha w_n) = 1, \quad (4)$$

$$\lim_{\alpha \rightarrow \infty} \limsup_{n \rightarrow \infty} d_n(t_{mix}^{(n)} + \alpha w_n) = 0. \quad (5)$$

For an introduction to Markov chains and mixing time see the highly recommended book by Levin, Peres and Wilmer (Levin et al., 2009), from which we borrow heavily.

2 The Erdős-Rényi random graph model

The easiest model in the class of rotationally symmetric and strictly unimodal distributions π on the hypercube arises when we have $\theta_k = \theta \in (0, 1]$ for all k , i.e. the acceptance probabilities for upward moves are constant across level sets. That is, $\pi(x) = \theta^{S(x)}(1 + \theta)^{-n}$. (The case where $\theta > 1$ would correspond to the mode of π being at $\underline{1} = (1, 1, \dots, 1)$ instead of $\underline{0} = (0, \dots, 0)$. By symmetry, this gives rise to nothing new, so we will assume $\theta \in (0, 1)$ henceforth. The case where $\theta = 1$ corresponds to π being the uniform distribution.)

If we have $n = \binom{\nu}{2}$ and identify the list of coordinates of the hypercube with the list of potential edges of a graph on ν vertices, then the hypercube represents the space of all possible (undirected) graphs on ν vertices. (A one indicates that a certain edge is present in the graph; a zero indicates that it is absent.) Since $\pi(x) = \left(\frac{\theta}{1+\theta}\right)^{S(x)} \left(1 - \frac{\theta}{1+\theta}\right)^{n-S(x)}$ for all $x \in \{0, 1\}^n$, this distribution π corresponds to the Erdős-Rényi random graph model with parameter $\frac{\theta}{1+\theta}$. This is the probability distribution on (undirected) graphs on ν vertices where each of the $n = \binom{\nu}{2}$ potential edges is present independently with probability $p := \frac{\theta}{1+\theta}$.

The non-lazy version of the random walk Metropolis chain for this model has cutoff at $\frac{1}{2(1+\theta)}n \log n$ with window size n . This was derived using Fourier analysis in Diaconis and Hanlon (1992, Theorem 2, page 104) together with Ross and Xu (1994, Theorem 4.2, page 824). Diaconis and Hanlon studied the projection $S(X_t)$ and explicitly calculated the eigenvalues and eigenvectors of the transition kernel. This establishes the lower bound. Ross and Xu showed that the Metropolis chain for this model on the hypercube can be viewed as a random walk on a hypergroup deformation of the hypercube. Fourier analysis of this random walk then leads to the result for the upper bound. Another proof using representation theory and Iwahori-Hecke algebras is given by Diaconis and

Ram (2000, Theorem 5.4, page 177). See also Diaconis and Saloff-Coste (2006, page 2117) for a discussion of these and some related results.

For the lazy version of this chain we therefore expect cutoff at $\frac{1}{1+\theta}n \log n$ with window size n . In this paper we will give an alternative proof of this result. This is done in the hope that a more probabilistic proof will be easier to generalize to less symmetric models, where Fourier analysis might be harder to apply. To be specific, we will prove the following result:

Theorem 2.1. *The lazy random walk Metropolis chain for $\pi(x) = \theta^{S(x)}(1+\theta)^{-n}$ on $\{0,1\}^n$ has cutoff at $\frac{1}{1+\theta}n \log n$ with a window of size n .*

Corollary 2.2. *Let $n := \binom{\nu}{2}$ and let $\pi(x) = p^{S(x)}(1-p)^{n-S(x)}$ be the Erdős-Rényi random graph model on ν vertices with parameter $p \in (0,1)$. The lazy random walk Metropolis chain for this model has cutoff at $\max\{p, 1-p\}n \log n$ with a window of size n .*

Proof of Corollary: As mentioned above, by relabeling states (switching zeros and ones) we may assume $p \in (0, 1/2]$. Let $\theta := \frac{p}{1-p}$. Since

$$p^{S(x)}(1-p)^{n-S(x)} = \left(\frac{\theta}{1+\theta}\right)^{S(x)} \left(1 - \frac{\theta}{1+\theta}\right)^{n-S(x)} = \theta^{S(x)}(1+\theta)^{-n}$$

and $\max\{p, 1-p\} = 1-p = \frac{1}{1+\theta}$, the result follows from the Theorem. \square

3 Lower bound

Our proof for the lower bound part of Theorem 2.1 mimics the one given in Levin et al. (2009, Proposition 7.13 on page 95) for the case where the stationary distribution π is uniform ($\theta = 1$). It is based on the method of distinguishing statistics, described for example in Levin et al. (2009). Their Proposition 7.8 on page 92 is this:

Proposition 3.1. *Let μ and ν be two probability distributions on \mathcal{X} , and let S be a real-valued function on \mathcal{X} . If*

$$|E_\mu(S) - E_\nu(S)| \geq r\sigma,$$

where $\sigma^2 = [\text{Var}_\mu(S) + \text{Var}_\nu(S)]/2$, then

$$\|\mu - \nu\|_{TV} \geq 1 - \frac{4}{4 + r^2}.$$

Here $E_\mu(S) := \sum_{x \in \mathcal{X}} S(x)\mu(x)$ denotes the expectation of S under μ , and likewise for ν . So if we can find a real function on the state space \mathcal{X} such that its expectations under $P^t(x, \cdot)$ and π are

still very different (on the scale of their average variance) after t steps of the chain, then we have demonstrated that $\|P^t(x, \cdot) - \pi\|$ must still be large.

A natural choice for the distinguishing statistic is the number of ones in a state, $S(x) := \sum_{i=1}^n x_i$ for $x \in \{0, 1\}^n$. Therefore we have to analyze the one-dimensional projection $S(X_t) =: V_t$ of our Markov chain (X_t) . As mentioned in the introduction, this is again a Markov chain whose transition probabilities satisfy $P(k, l) = P(x, \cdot)S^{-1}(l) = P(x, [y])$ for any x, y with $S(x) = k$ and $S(y) = l$. As before, $[y] = \{z \in \{0, 1\}^n : S(z) = S(y)\}$ denotes the equivalence class of all states with the same number of ones as y . (Because of their different domains it should not lead to confusion that we are using the same notation $P(\cdot, \cdot)$ for the transition probabilities of the original chain and the projected chain.)

Similarly, the stationary distribution π_S of (V_t) is the push-forward $\pi_S := \pi S^{-1}$ of π under S . This entails $\pi_S(k) = \binom{n}{k} \theta^k (1 + \theta)^{-n} = \binom{n}{k} \left(\frac{\theta}{1+\theta}\right)^k \left(1 - \frac{\theta}{1+\theta}\right)^{n-k} = \text{Binomial}\left(n, \frac{\theta}{1+\theta}\right)(k)$.

Therefore we get the expectation and variance of $V \sim \pi_S$ as

$$E_{\pi_S} V = \frac{n \theta}{1 + \theta}, \quad \text{Var}_{\pi_S} V = \frac{n \theta}{(1 + \theta)^2}. \quad (6)$$

The chain (V_t) is a birth and death chain on $\{0, 1, \dots, n\}$ with transition probabilities

$$\begin{aligned} p_k &:= P(V_{t+1} = k + 1 | V_t = k) = \left(1 - \frac{k}{n}\right) \frac{\theta}{2}, \\ r_k &:= P(V_{t+1} = k | V_t = k) = \frac{1}{2} + \left(1 - \frac{k}{n}\right) \frac{1 - \theta}{2}, \\ q_k &:= P(V_{t+1} = k - 1 | V_t = k) = \frac{k}{2n}. \end{aligned} \quad (7)$$

We begin by calculating its expectation after t steps starting from k . Note that for all t we get

$$V_{t+1} - V_t = \begin{cases} 1 & \text{with probability } \left(1 - \frac{V_t}{n}\right) \frac{\theta}{2}, \\ -1 & \text{with probability } \frac{V_t}{2n}, \end{cases}$$

and $V_{t+1} - V_t = 0$ otherwise. So

$$\mathbf{E}[V_{t+1} - V_t | V_t] = \left(1 - \frac{V_t}{n}\right) \frac{\theta}{2} - \frac{V_t}{2n} = \frac{\theta}{2} - V_t \frac{1 + \theta}{2n},$$

and therefore

$$\mathbf{E}[V_{t+1} | V_t] = \frac{\theta}{2} - V_t \frac{1 + \theta}{2n} + E[V_t | V_t] = \frac{\theta}{2} + \left(1 - \frac{1 + \theta}{2n}\right) V_t.$$

By taking expectation \mathbf{E}_k with respect to the starting state k we get for all t, k that

$$\mathbf{E}_k(V_{t+1}) = \frac{\theta}{2} + \left(1 - \frac{1 + \theta}{2n}\right) \mathbf{E}_k(V_t). \quad (8)$$

This leads to the following result:

Proposition 3.2. *The projected chain $V_t := S(X_t)$ of our Metropolis chain (X_t) has for all $k = 0, 1, \dots, n$ and all $t \in \mathbb{N}$*

$$\mathbf{E}_k(V_t) = \frac{n\theta}{1+\theta} (1 - \gamma^t) + k\gamma^t, \quad (9)$$

where $\gamma := 1 - \frac{1+\theta}{2n}$.

Proof: By induction on t . Fix any starting state $k \in \{0, 1, \dots, n\}$. The statement is true for $t = 0$ since $E_k(V_0) = k$. Now suppose it is true for t . Then by (8) and the induction hypothesis we get that

$$\begin{aligned} E_k(V_{t+1}) &= \frac{\theta}{2} + \gamma \left[\gamma^t \left(k - \frac{n\theta}{1+\theta} \right) + \frac{n\theta}{1+\theta} \right] \\ &= \frac{\theta}{2} + \gamma^{t+1} \left(k - \frac{n\theta}{1+\theta} \right) + \left(\frac{n\theta}{1+\theta} - \frac{\theta}{2} \right) \\ &= \gamma^{t+1} \left[k - \frac{n\theta}{1+\theta} \right] + \frac{n\theta}{1+\theta}, \end{aligned}$$

so it is also true for $t + 1$. \square

Remark 3.3: Note that for $u := u_{n,\theta} := \frac{1}{1+\theta} n \log n$ we get

$$\gamma^u \sim n^{-1/2},$$

by which we mean that $\lim_{n \rightarrow \infty} \frac{\gamma^u}{n^{-1/2}} = 1$.

Proof: Using $1 - x \leq e^{-x}$ for $x \in \mathbb{R}$ we get

$$\begin{aligned} \lim_n \frac{\gamma^u}{n^{-1/2}} &= \lim_n \sqrt{n} \left(1 - \frac{1+\theta}{2n} \right)^{\frac{1}{1+\theta} n \log n} \\ &\leq \lim_n \sqrt{n} \exp \left\{ -\frac{1+\theta}{2n} \frac{1}{1+\theta} n \log n \right\} \\ &= \lim_n \sqrt{n} \exp \left\{ -\frac{1}{2} \log n \right\} \\ &= \lim_n \sqrt{n} n^{-1/2} \\ &= 1. \end{aligned}$$

For the other direction we expand the natural logarithm about one to get

$$\begin{aligned} -\log \left(1 - \frac{1+\theta}{2n} \right) &= \sum_{k=1}^{\infty} \frac{1}{k} \left(\frac{1+\theta}{2n} \right)^k \\ &\leq \frac{1+\theta}{2n} + \frac{1}{2} \left(\frac{1+\theta}{2n} \right)^2 + \frac{1}{3} \sum_{k=3}^{\infty} \left(\frac{1+\theta}{2n} \right)^k \\ &= \frac{1+\theta}{2n} + O(1/n^2). \end{aligned}$$

This means that for some positive constant C we get

$$\begin{aligned}
\lim_n \frac{\gamma^u}{n^{-1/2}} &= \lim_n \sqrt{n} \exp \left\{ \frac{1}{1+\theta} n \log(n) \log \left(1 - \frac{1+\theta}{2n} \right) \right\} \\
&\geq \lim_n \sqrt{n} \exp \left\{ -\frac{1}{1+\theta} n \log(n) \left[\frac{1+\theta}{2n} + O(1/n^2) \right] \right\} \\
&\geq \lim_n \sqrt{n} \exp \left\{ -\frac{1}{2} \log(n) \right\} \exp \left\{ -\frac{1}{1+\theta} \frac{C}{n} \log n \right\} \\
&= \lim_n \sqrt{n} n^{-1/2} \\
&= 1. \quad \square
\end{aligned}$$

It remains to calculate the variance of V_t . Since we want a lower bound on

$$d(t) = \sup_x \|P^t(x, \cdot) - \pi\| \geq \|P^t(\underline{1}, \cdot) - \pi\|,$$

it's enough to consider the starting state $\underline{1} = (1, \dots, 1)$ of all ones. For this, first note that we can run the chain X_t in the following way. Given we are at state X_t at time t :

- Pick a coordinate $i \in [n]$ uniformly at random.
- Draw $U_t \sim \text{Uniform}[0, 1]$, independent of everything else.
- Set $X_{t+1}^{(j)} := X_t^{(j)}$ for $j \neq i$, and set the i^{th} coordinate of X_{t+1} to

$$X_{t+1}^{(i)} := \begin{cases} 1 & : 0 \leq U_t \leq \frac{\theta}{2}, \\ X_t^{(i)} & : \frac{\theta}{2} < U_t \leq \frac{1}{2}, \\ 0 & : \frac{1}{2} < U_t \leq 1. \end{cases}$$

Now say that a coordinate j has been *refreshed* by time t , if coordinate j was selected at some time $s < t$ and $U_s \notin (\frac{\theta}{2}, \frac{1}{2}]$. Let R_t be the number of coordinates not refreshed by time t . We can study the expectation and variance of R_t with a natural modification of the classical coupon collector problem. This leads to the following result, analogous to Lemma 7.12 on page 94 in Levin et al. (2009):

Proposition 3.4. *Consider the coupon collector problem with n distinct coupon types, where at each trial, with probability $\frac{1-\theta}{2}$ we get no coupon, and with probability $1 - \frac{1-\theta}{2}$ we get a coupon chosen (independently and) uniformly at random. Let $I_j(t)$ be the indicator of the event that the j^{th} coupon has not been collected by time t . Let $R_t := \sum_{i=1}^n I_i(t)$ be the number of coupon types not collected by time t . The random variables $I_j(t)$ are negatively correlated, and letting $\gamma^t := \left(1 - \frac{1+\theta}{2n}\right)^t$, we get for $t \geq 0$ that*

$$\begin{aligned}
E(R_t) &= n\gamma^t, \\
\text{Var}(R_t) &\leq n\gamma^t(1 - \gamma^t) \leq n\gamma^t.
\end{aligned}$$

Proof: By definition of $I_j(t)$ we get

$$E[I_j(t)] = P\{\text{No coupon of type } j \text{ in trials } 1, \dots, t\} = \left(1 - \frac{1+\theta}{2n}\right)^t = \gamma^t$$

and

$$\text{Var}[I_j(t)] = E[(I_j(t))^2] - (E[I_j(t)])^2 = \gamma^t - \gamma^{2t} = \gamma^t(1 - \gamma^t).$$

Similarly, for $j \neq k$ we get

$$\begin{aligned} E[I_j(t) I_k(t)] &= P\{\text{No coupon of type } j \text{ or } k \text{ in trials } 1, \dots, t\} \\ &= \left(1 - \frac{1+\theta}{2} \frac{2}{n}\right)^t \\ &= \left(1 - \frac{1+\theta}{n}\right)^t, \end{aligned}$$

so

$$\text{Cov}[I_j(t), I_k(t)] = E[I_j(t) I_k(t)] - E[I_j(t)] E[I_k(t)] = \left(1 - \frac{1+\theta}{n}\right)^t - \left(1 - \frac{1+\theta}{2n}\right)^{2t} \leq 0.$$

Therefore

$$E[R_t] = \sum_{j=1}^n E[I_j(t)] = n\gamma^t$$

and

$$\text{Var}[R_t] = \sum_{j=1}^n \text{Var}[I_j(t)] + \sum_{j \neq k} \text{Cov}[I_j(t), I_k(t)] \leq n\gamma^t(1 - \gamma^t). \quad \square$$

If we start the chain at $X_0 = \underline{1}$, then the conditional distribution of $V_t = S(X_t)$ given $R_t = r$ is the same as that of $r + B$, where $B \sim \text{Binomial}(n - r, \frac{\theta}{1+\theta})$. Therefore,

$$E_{\underline{1}}[V_t | R_t] = R_t + (n - R_t) \frac{\theta}{1+\theta} = \frac{R_t + n\theta}{1+\theta},$$

so by taking expectation we get

$$E_{\underline{1}}[V_t] = \frac{E[R_t] + n\theta}{1+\theta} = \frac{n\gamma^t + n\theta}{1+\theta} \left(= \frac{n\theta}{1+\theta}(1 - \gamma^t) + n\gamma^t \right),$$

confirming our result for general starting states k from above for the special case $k = n$. Furthermore, since

$$\text{Var}_{\underline{1}}[V_t] = \text{Var} [E_{\underline{1}}(V_t | R_t)] + E [\text{Var}_{\underline{1}}(V_t | R_t)],$$

we get

$$\begin{aligned}
Var_{\underline{1}}[V_t] &= Var \left[\frac{R_t + n\theta}{1 + \theta} \right] + E \left[Var(\text{Binomial}(n - R_t, \frac{\theta}{1 + \theta})) \right] \\
&= \frac{1}{(1 + \theta)^2} Var[R_t] + \frac{1}{(1 + \theta)^2} (n - E[R_t]) \theta \\
&\leq \frac{1}{(1 + \theta)^2} [n\gamma^t + (n - n\gamma^t)\theta] \\
&\leq \frac{n}{(1 + \theta)^2}.
\end{aligned}$$

To apply Proposition 3.1, observe that

$$\sigma^2 := \frac{Var_{P^t(\underline{1}, \cdot)} V + Var_{\pi} V}{2} \leq \max\{Var_{\underline{1}} V_t, Var_{\pi_S} V\} \leq \frac{n}{(1 + \theta)^2}.$$

So for $t := t_{n, \alpha} := \frac{1}{1 + \theta} n \log n - \alpha n$ and $\gamma^t = (1 - \frac{1 + \theta}{2n})^t$ we get for any fixed $\varepsilon > 0$ and large n that

$$\begin{aligned}
|E_{\underline{1}} V_t - E_{\pi} V| &= \left| \frac{n(\theta + \gamma^t)}{1 + \theta} - \frac{n\theta}{1 + \theta} \right| \\
&= \frac{n\gamma^t}{1 + \theta} \\
&= \frac{n}{1 + \theta} \left(1 - \frac{1 + \theta}{2n} \right)^t \\
&\geq \sigma \sqrt{n} \left(1 - \frac{1 + \theta}{2n} \right)^{n(\frac{1}{1 + \theta} \log n - \alpha)} \\
&\geq \sigma \sqrt{n} (1 - \varepsilon) \exp \left\{ -\frac{1 + \theta}{2} \left(\frac{1}{1 + \theta} \log n - \alpha \right) \right\} \\
&= \sigma (1 - \varepsilon) \exp \left\{ \alpha \frac{1 + \theta}{2} \right\} \\
&=: \sigma r_{\alpha}.
\end{aligned}$$

By Proposition 3.1, this means $d(t) \geq \|P^t(\underline{1}, \cdot) - \pi\| \geq 1 - \frac{4}{4 + r_{\alpha}^2}$. Therefore

$$\lim_{\alpha \rightarrow \infty} \liminf_{n \rightarrow \infty} d(t_{n, \alpha}) \geq \lim_{\alpha \rightarrow \infty} 1 - \frac{4}{4 + r_{\alpha}^2} = 1.$$

This finishes the proof of the lower bound part of Theorem 2.1. \square

4 Lower bound, alternative proof

The previous proof for the lower bound part of Theorem 2.1 might be hard to generalize to cases where θ_k is not constant in k . (Recall that $\theta_k := \frac{\pi(x)}{\pi(y)}$ for any x, y such that $S(x) = k + 1, S(y) = k$.) Therefore, we give here an alternative proof for a slightly weaker result.

Proposition 4.1. *Let (X_t) be the lazy random walk Metropolis chain for $\pi(x) = \theta^{S(x)}(1 + \theta)^{-n}$ on $\{0, 1\}^n$. Then*

$$\lim_{n \rightarrow \infty} d_n(C \frac{1}{1 + \theta} n \log n) = 1 \quad \text{for all } C < 1. \quad (10)$$

Note that (10) establishes the lower bound part of the result that this chain has cutoff at $\frac{1}{1 + \theta} n \log n$. It does not give the window size though.

For this proof we use a modified version of the method of distinguishing statistics (see Proposition 3.1) that avoids the need to calculate or approximate the variance of the statistic S under $P^t(x, \cdot)$. Instead, we use the fact that our chain only makes *local moves* to argue that $S(X_t) = V_t$ must be concentrated about its mean under $P^t(x, \cdot)$. To do this, we use Azuma's inequality:

Theorem 4.2 (Azuma's Inequality). *: Let $(Y_i)_{i=0,1,\dots}$ be a martingale with bounded differences, i.e.*

$$a_i \leq Y_i - Y_{i-1} \leq b_i \quad \text{for some } a_i, b_i \in \mathbb{R} \text{ and all } i \in \mathbb{N}.$$

Then for any $t \in \mathbb{N}$ and $s > 0$ we get

$$\begin{aligned} P\{Y_t - Y_0 \geq s\} &\leq e^{-2s^2/c} & \text{and} \\ P\{Y_t - Y_0 \leq -s\} &\leq e^{-2s^2/c}, \end{aligned}$$

where $c := \sum_{i=1}^t (b_i - a_i)^2$.

For a proof see for example Dubhashi & Panconesi (2009, Theorem 5.2, page 67) or Ross (1996, Theorem 6.3.3, page 307). Note that both inequalities are strict unless Y_t is constant a.s., since this is true for Markov's inequality (applied to e^{Y_t}), on which the proof is based.

Now we can't apply this result directly in our situation, since $V_i = S(X_i)$ is not a martingale. However, by the tower property of conditional expectation, $Y_i := E[V_t | V_{0:i}]$ for $i = 0, 1, \dots, t$ gives a martingale with respect to V_i . Here we write $V_{k:l} := (V_i)_{k \leq i \leq l}$. This immediately gives the following result, sometimes called the method of averaged bounded differences (e.g. in Dubhashi & Panconesi, 2009, Theorem 5.3 on page 68).

Proposition 4.3. *Let $V_0 = v \in \mathbb{R}$ and $t \in \mathbb{N}$ be fixed and let V_1, \dots, V_t be a sequence of real-valued random variables with averaged bounded differences, i.e.*

$$|E[V_t | V_{0:i}] - E[V_t | V_{0:(i-1)}]| \leq c_i$$

for some $c_i \in \mathbb{R}$ and all $i = 1, \dots, t$. Then for any $s > 0$ we get

$$\begin{aligned} P_v\{V_t \geq E_v V_t + s\} &\leq e^{-2s^2/c} & \text{and} \\ P_v\{V_t \leq E_v V_t - s\} &\leq e^{-2s^2/c}, \end{aligned}$$

where $c := 4 \sum_{i=1}^t c_i^2$.

Proof: Fix $t \in \mathbb{N}$ and consider $Y_i := E[V_t | V_{0:i}]$ for $i = 0, 1, \dots, t$. This gives a martingale with respect to (V_i) . Furthermore, $Y_0 = E[V_t | V_0] = E_v V_t$ and $Y_t = V_t$. By assumption, the martingale (Y_i) has bounded differences

$$-c_i \leq Y_i - Y_{i-1} \leq c_i \text{ for all } i \in [t].$$

Applying Azuma's inequality to (Y_i) finishes the proof. \square

Now we apply this result to our Markov chain $V_t = S(X_t)$ to show that it is concentrated about its mean. This will follow from Proposition 4.3 where we use the result from Proposition 3.2 to show that (V_t) has averaged bounded differences.

Proposition 4.4. *Let (X_t) be the lazy random walk Metropolis chain on $\{0, 1\}^n$ for $\pi(x) = \theta^{S(x)}(1 + \theta)^{-n}$. Let $V_t := S(X_t)$ and also $X_0 = x$, so $v := V_0 = S(x)$. Then for any $s > 0$ we get*

$$\begin{aligned} P_v\{V_t \geq E_v V_t + s\} &\leq e^{-2s^2/(9t)} \quad \text{and} \\ P_v\{V_t \leq E_v V_t - s\} &\leq e^{-2s^2/(9t)}. \end{aligned}$$

Proof: Let $\gamma := 1 - \frac{1+\theta}{2n}$ and fix any $t \in \mathbb{N}$. Since (V_i) is a Markov chain, we get from Proposition 3.2 and the Markov property that for any $i = 0, 1, \dots, t$

$$E[V_t | V_{0:i}] = E_{V_i}[V_{t-i}] = \gamma^{t-i} \left[V_i - \frac{n\theta}{1+\theta} \right] + \frac{n\theta}{1+\theta}.$$

Therefore for any $i = 0, 1, \dots, t$ we get

$$\begin{aligned} E[V_t | V_{0:i}] - E[V_t | V_{0:(i-1)}] &= \gamma^{t-i} \left[V_i - \frac{n\theta}{1+\theta} \right] - \gamma^{t-(i-1)} \left[V_{i-1} - \frac{n\theta}{1+\theta} \right] \\ &= \gamma^{t-i} \left[V_i - V_{i-1} + (1-\gamma)V_{i-1} - (1-\gamma)\frac{n\theta}{1+\theta} \right] \\ &= \gamma^{t-i} \left[V_i - V_{i-1} + \frac{1+\theta}{2n}V_{i-1} - \frac{\theta}{2} \right], \end{aligned}$$

and so

$$|E[V_t | V_{0:i}] - E[V_t | V_{0:(i-1)}]| \leq \gamma^{t-i} \frac{3}{2} \leq \frac{3}{2},$$

since we have $V_i \in [0, n]$ and $|V_i - V_{i-1}| \leq 1$, and also $\theta \in (0, 1]$. Applying Proposition 4.3 with $c_i := \frac{3}{2}$ now gives the result. \square

Now we use the concentration result for (V_t) from Proposition 4.4 to establish a lower bound on the mixing time for this Markov chain.

Proof of Proposition 4.1: Fix $\delta \in (0, \frac{1}{2})$ and set $t := t_{n,\delta} := \frac{1}{1+\delta} \frac{1}{1+\theta} n \log n$. To establish (10) from Proposition 4.1, we need to show

$$\lim_{n \rightarrow \infty} d(t_{n,\delta}) = 1. \quad (11)$$

(Note that since $d(t)$ is decreasing in t , it is enough to prove (10) for *large* $c < 1$, so our restriction $\delta < \frac{1}{2}$ is not problematic.) To this end, denote with $V_t := S(X_t)$ the projection of the chain under S , started at $X_0 = \underline{1}$, so that $V_0 = n$ and note that for any $r \in [0, n]$ we get

$$\begin{aligned} d(t) &\geq \|P^t(\underline{1}, \cdot) - \pi\| \\ &= \sup_{A \subset \{0,1\}^n} P^t(\underline{1}, A) - \pi(A) \\ &\geq \sup_{L \subset [n]} P^t(\underline{1}, S^{-1}(L)) - \pi(S^{-1}(L)) \\ &= \sup_{L \subset [n]} P^t(n, L) - \pi_S(L) \\ &\geq P_n\{V_t \geq r\} - P\{B_n \geq r\} \\ &= 1 - P_n\{V_t < r\} - P\{B_n \geq r\}. \end{aligned}$$

Here the second inequality comes from noting that the right supremum is over a subset of events from the left supremum. (Therefore, total variation distance can only decrease after projecting down.) Furthermore, B_n stands for a random variable with distribution $\pi_S = \text{Binomial}(n, \frac{\theta}{1+\theta})$ and we used the event $L := \{\lceil r \rceil, \lceil r \rceil + 1, \dots, n\}$ for the last inequality. Our task is to find $r \in [0, n]$ such that

$$\lim_{n \rightarrow \infty} P_n\{V_t < r\} = 0 \quad \text{and} \quad (12)$$

$$\lim_{n \rightarrow \infty} P\{B_n \geq r\} = 0. \quad (13)$$

To this end let $\rho := \frac{\delta}{4(1+\delta)}$ and $c := c_{n,\delta} := 1 - n^{-\rho}$, so $c \in (0, 1)$, and define a convex combination of the means of the two random variables V_t and B_n by

$$r := r_{n,\delta} := c E_n V_t + (1 - c) \frac{n\theta}{1 + \theta} = \frac{n\theta}{1 + \theta} + c\gamma^t n \frac{1}{1 + \theta}.$$

Here we used $\gamma := 1 - \frac{1+\theta}{2n}$ and Proposition 3.2 for the last equality. Define

$$\begin{aligned} s_1 &:= E_n V_t - r = (1 - c) \left[E_n V_t - \frac{n\theta}{1 + \theta} \right] = n^{-\rho} \gamma^t n \frac{1}{1 + \theta}, \\ s_2 &:= r - \frac{n\theta}{1 + \theta} = c \left[E_n V_t - \frac{n\theta}{1 + \theta} \right] = (1 - n^{-\rho}) \gamma^t n \frac{1}{1 + \theta}. \end{aligned}$$

To establish (12) we use the concentration result for V_t from Proposition 4.4. This shows

$$P_n\{V_t < r\} = P_n\{V_t < E_n V_t - s_1\} \leq \exp\left\{-\frac{2s_1^2}{9t}\right\}.$$

Now fix any $\varepsilon \in (0, 1)$ and we get for large n that

$$\begin{aligned}
\frac{s_1^2}{t} &= \frac{n^{-2\rho} \gamma^{2t} n^2}{t(1+\theta)^2} \\
&= \frac{n^{2-2\rho} \left(1 - \frac{1+\theta}{2n}\right)^{n \frac{2}{1+\theta} \frac{1}{1+\delta} \log n}}{\frac{1+\theta}{1+\delta} n \log n} \\
&\geq \frac{n^{1-2\rho} (1-\varepsilon) \exp\left\{-\frac{1+\theta}{2} \frac{2}{1+\theta} \frac{1}{1+\delta} \log n\right\}}{\frac{1+\theta}{1+\delta} \log n} \\
&= \frac{(1+\delta)(1-\varepsilon)}{1+\theta} \frac{n^{1-2\rho-\frac{1}{1+\delta}}}{\log n}.
\end{aligned} \tag{14}$$

Here the inequality comes from $(1 - \frac{x}{n})^n \rightarrow e^{-x}$ as n goes to infinity for any $x \in \mathbb{R}$. Now the right hand side of this expression goes to infinity as n goes to infinity because the left ratio is a positive constant and the right ratio goes to infinity as n goes to infinity since $1 - 2\rho - \frac{1}{1+\delta} > 0$. This proves (12).

To establish (13), we use Chebyshev's inequality for the $\text{Binomial}(n, \frac{\theta}{1+\theta})$ random variable B_n . We get for any fixed $\varepsilon \in (0, 1)$ and large n that

$$\begin{aligned}
P\{B_n \geq r\} &= P\{B_n - \frac{n\theta}{1+\theta} \geq s_2\} \\
&\leq \frac{\text{Var} B_n}{s_2^2} \\
&= \frac{\frac{n\theta}{(1+\theta)^2}}{c^2 \gamma^{2t} n^2 \left(\frac{1}{1+\theta}\right)^2} \\
&= \frac{\theta}{c^2 n \left(1 - \frac{1+\theta}{2n}\right)^{n \left[\frac{2}{1+\theta} \frac{1}{1+\delta} \log n\right]}} \\
&\leq \frac{\theta}{c^2 n (1-\varepsilon) \exp\left\{-\frac{1+\theta}{2} \frac{2}{1+\theta} \frac{1}{1+\delta} \log n\right\}} \\
&= \frac{\theta}{c^2 (1-\varepsilon) n^{1-\frac{1}{1+\delta}}}.
\end{aligned}$$

The last inequality again comes from $(1 - \frac{x}{n})^n \rightarrow e^{-x}$ as n goes to infinity for any $x \in \mathbb{R}$. Now the right hand side of this expression goes to zero as n goes to infinity, since $c = c_{n,\delta}$ goes to one as n goes to infinity, and $1 - \frac{1}{1+\delta} = \frac{\delta}{1+\delta} > 0$. This proves (13) and therefore (11), finishing the proof of Theorem 2.1, lower bound (10). \square

Remark 4.5: Note that our concentration result for V_t from Azuma's inequality is not strong enough to establish the window size n of the cutoff. To see this, note that the smallest r that

ensures that $P\{B_n \geq r\}$ goes to zero as α goes to infinity would be of the form $r = \frac{n\theta}{1+\theta} + \alpha\sqrt{n}$, since $EB_n = \frac{n\theta}{1+\theta}$ and $Var B_n = O(n)$. However, in that case for $t_{n,\alpha} := \frac{1}{1+\theta}n \log n - \alpha n$, Azuma's inequality only gives a trivial upper bound of one for $\lim_{\alpha \rightarrow \infty} \lim_{n \rightarrow \infty} P_n\{V_{t_{n,\alpha}} < r\}$.

5 Upper bound

To establish the upper bound part of Theorem 2.1, we use a two-stage coupling procedure similar to the one used in Levin et al. (2009, Theorem 18.3, page 251) for the case where π is uniform ($\theta = 1$). The first step there is to show that upper bounding $d(t) = \sup_x \|P^t(x, \cdot) - \pi\|$ can be reduced to the the problem of upper bounding $d_S(t) = \sup_k \|P^t(k, \cdot) - \pi_S\|$, i.e. to show that it is enough to analyze the mixing time of the one-dimensional projection $S(X_t) \in \{0, 1, \dots, n\}$ of $X_t \in \{0, 1\}^n$.

This is possible for $\theta = 1$ since by symmetry, the total variation distance to stationarity is the same for all starting states:

$$\|P^t(x, \cdot) - \pi\| = \|P^t(y, \cdot) - \pi\| \quad \text{for all } x, y \in \{0, 1\}^n.$$

Therefore it is enough to consider the starting state $X_0 = \underline{1} = (1, 1, \dots, 1)$ of all ones. But for this starting state the transition probabilities are constant on level sets $L(k) := \{z \in \{0, 1\}^n : S(z) = k\}$, where $k = 0, 1, \dots, n$. By induction on t , we get this also for the t -step transition probabilities:

$$P^t(\underline{1}, y) = P^t(\underline{1}, z) \quad \text{whenever } S(y) = S(z).$$

Since π is also constant on level sets, this entails

$$\begin{aligned} \|P^t(\underline{1}, \cdot) - \pi\| &= \frac{1}{2} \sum_{l=0}^n \sum_{z: S(z)=l} |P^t(\underline{1}, z) - \pi(z)| \\ &= \frac{1}{2} \sum_{l=0}^n \left| \sum_{z: S(z)=l} P^t(\underline{1}, z) - \pi(z) \right| \\ &= \frac{1}{2} \sum_{l=0}^n |P^t(n, l) - \pi_S(l)| \\ &= \|P^t(n, \cdot) - \pi_S\|, \end{aligned}$$

where we can move the absolute values outside the (inner) sum in the second equality because all the terms in the sum are equal. So total variation distance stays the same under the one-dimensional projection S if we start from $X_0 = \underline{1}$ (or $\underline{0}$).

When $\theta \in (0, 1)$, the t -step transition probabilities are still constant on level sets if we start the chain at $X_0 = \underline{1}$ (or $\underline{0}$). However, total variation distance is not necessarily the same for each

starting state in this case. Also, we have been unable to show that $X_0 = \underline{1}$ (or $\underline{0}$) is a worst starting state, i.e. we couldn't prove

$$d(t) = \sup_x \|P^t(x, \cdot) - \pi\| = \max\{\|P^t(\underline{1}, \cdot) - \pi\|, \|P^t(\underline{0}, \cdot) - \pi\|\},$$

so we don't know whether bounding the mixing time for $\theta \in (0, 1)$ can be reduced to a one-dimensional problem in the same way as when π is uniform.

However, we have been able to get an upper bound on the mixing time using a *two*-dimensional projection that depends on the starting state $X_0 = x$. For this, consider $Z := Z_x : \{0, 1\}^n \rightarrow \{0, 1, \dots, n\} \times \{0, 1, \dots, n\}$,

$$X_t \mapsto Z_x(X_t) := Z_t := (S(X_t), d(x, X_t)),$$

where d is graph distance on the hypercube (i.e. $d(x, y)$ is the number of coordinates where x and y disagree) and $x = X_0$ is the starting state of the chain. In words, we project down to the number of ones, $S(X_t)$, and the distance to the starting state, $d(X_t, X_0)$. This two-dimensional chain is very similar to the two-coordinate chain used in Levin, Luczak and Peres (2010, section 3) for the related model of Glauber dynamics for the mean-field Ising model.

Our proof proceeds by showing that (Z_t) is a (two-dimensional) birth and death chain with the same total variation distance to its stationary distribution as the original chain X_t . Bounding this distance is then achieved by a two-stage coupling procedure. In the first stage (*drift*-regime) we use an independence coupling of two versions of the chain that brings them close together after $\frac{1}{1+\theta}n \log n$ steps due to the drift towards the mean in this birth and death chain. In the second stage (*entropy*-regime) we use a coupling with a certain Ornstein-type coupling of two lazy simple random walks on \mathbb{Z}^2 to ensure that the two versions of the chain coalesce after an additional αn steps.

We begin by establishing some properties of the two-dimensional projection $Z_t := Z_x(X_t)$ of X_t , for which we need some more notation. Fix $x, z \in \{0, 1\}^n$ and let $S(x) = k, S(z) = l, d(x, z) = l'$. Define

$$\begin{aligned} G &:= G(x, z) := \{i \in [n] : x^{(i)} = 0, z^{(i)} = 0\}, \\ N &:= N(x, z) := \{i \in [n] : x^{(i)} = 0, z^{(i)} = 1\}, \\ E &:= E(x, z) := \{i \in [n] : x^{(i)} = 1, z^{(i)} = 0\}, \\ F &:= F(x, z) := \{i \in [n] : x^{(i)} = 1, z^{(i)} = 1\}. \end{aligned}$$

When clear from the context, we might suppress the dependence of G, N, E, F on x, z in the notation. For the number of elements in these four sets we get

$$\#G = n - \frac{l + l' + k}{2},$$

$$\begin{aligned}
\#N &= \frac{l + l' - k}{2}, \\
\#E &= \frac{l' + k - l}{2}, \\
\#F &= \frac{l - (l' - k)}{2}.
\end{aligned} \tag{15}$$

This follows from $\#N + \#E = l'$, $\#N + \#F = l$, $\#E + \#F = k$, $\#G + \#N = n - k$, e.g. by starting with the observation

$$l' - l + \#F = l' - \#N = \#E = k - \#F,$$

which gives the last equality $2\#F = l - (l' - k)$ above. The remaining equalities then follow. This is probably best understood by looking at an example. Suppose

$$\begin{aligned}
x &= 00000\,000\,11\,1, \\
z &= 00000\,111\,00\,1.
\end{aligned}$$

Then $n = 11$, $S(x) = k = 3$, $S(z) = l = 4$, $d(x, z) = 5$ and $\#G = 5$, $\#N = 3$, $\#E = 2$, $\#F = 1$. Thus, we have a one-to-one correspondence between (n, k, l, l') and $(\#G, \#N, \#E, \#F)$. Note that the formulas for $\#G, \#N, \#E, \#F$ above all give integers because $l, l' - k, l' + k$ always have the same parity.

Proposition 5.1. *For each $x \in \{0, 1\}^n$ the t -step transition probabilities $P^t(x, \cdot)$ are constant on sets*

$$L(l) \cap D(x, l'), \quad \text{for } l, l' \in \{0, 1, \dots, n\},$$

where $L(l) := \{z \in \{0, 1\}^n : S(z) = l\}$ and $D(x, l') := \{z \in \{0, 1\}^n : d(x, z) = l'\}$.

Proof: Fix $x \in \{0, 1\}^n$ and let $k := S(x)$. We use induction on t . For $t = 1$, the only nontrivial cases to check are $l' = 1$ and $l - k = \pm 1$. For any $y, z \in L(l) \cap D(x, l')$ we get

$$\begin{aligned}
P(x, y) &= \frac{1}{2n}\theta = P(x, z) \quad \text{if } l = k + 1 \quad \text{and} \\
P(x, y) &= \frac{1}{2n} = P(x, z) \quad \text{if } l = k - 1,
\end{aligned}$$

so the statement is true for $t = 1$. Now suppose it's true for t and fix $l, l' \in \{0, 1, \dots, n\}$ as well as $y, z \in L(l) \cap D(x, l')$. Recall the definition of $G(x, z), N(x, z), E(x, z), F(x, z)$, etc. for these x, y, z from above and define

$$\begin{aligned}
\bar{G}(x, z) &= \{v \in \{0, 1\}^n : \exists i \in G(x, z) \text{ such that } v^{(i)} = z^{(i)} + 1, v^{(j)} = z^{(j)} \forall j \neq i\}, \\
\bar{N}(x, z) &= \{v \in \{0, 1\}^n : \exists i \in N(x, z) \text{ such that } v^{(i)} = z^{(i)} - 1, v^{(j)} = z^{(j)} \forall j \neq i\}, \\
\bar{E}(x, z) &= \{v \in \{0, 1\}^n : \exists i \in E(x, z) \text{ such that } v^{(i)} = z^{(i)} + 1, v^{(j)} = z^{(j)} \forall j \neq i\}, \\
\bar{F}(x, z) &= \{v \in \{0, 1\}^n : \exists i \in F(x, z) \text{ such that } v^{(i)} = z^{(i)} - 1, v^{(j)} = z^{(j)} \forall j \neq i\}.
\end{aligned}$$

Similarly for y . For example, $\bar{G}(x,z)$ is the set of v 's that agree with z everywhere except for one coordinate in $G(x,z)$, i.e. a coordinate where both x and z are zero. By assumption, y and z are in the same set $L(l) \cap D(x, l')$. So clearly, $\#\bar{G}(x,y) = \#G(x,y) = \#G(x,z) = \#\bar{G}(x,z)$. Furthermore, we get

$$\bar{G}(x,y), \bar{G}(x,z) \subset L(l+1) \cap D(x, l'+1).$$

This means that if $s \in \bar{G}(x,y)$, $s' \in \bar{G}(x,z)$, then s and s' are in the same two-dimensional level set $L(l+1) \cap D(x, l'+1)$. This implies $P(s,y) = P(s',z)$ by the transition rule for the Markov chain, and it further implies $P^t(x,s) = P^t(x,s')$ by the induction hypothesis. Similarly for N, E, F .

Therefore we get

$$\begin{aligned} P^{t+1}(x,z) &= P^t(x,z)P(z,z) + \sum_{s' \in \bar{G}(x,z)} P^t(x,s')P(s',z) + \sum_{s' \in \bar{N}(x,z)} P^t(x,s')P(s',z) + \\ &\quad + \sum_{s' \in \bar{E}(x,z)} P^t(x,s')P(s',z) + \sum_{s' \in \bar{F}(x,z)} P^t(x,s')P(s',z) \\ &= P^t(x,y)P(y,y) + \sum_{s \in \bar{G}(x,y)} P^t(x,s)P(s,y) + \sum_{s \in \bar{N}(x,y)} P^t(x,s)P(s,y) + \\ &\quad + \sum_{s \in \bar{E}(x,y)} P^t(x,s)P(s,y) + \sum_{s \in \bar{F}(x,y)} P^t(x,s)P(s,y) \\ &= P^{t+1}(x,y). \end{aligned}$$

So the statement is also true for $t+1$ and we are done. \square

Corollary 5.2. *The projection $Z_t := Z_x(X_t) := (S(X_t), d(x, X_t))$ is Markov and we get*

$$\|P^t(x, \cdot) - \pi\| = \|\mathcal{D}_k(Z_t) - \pi_{Z_x}\|,$$

where $x = X_0$ and $k = S(x)$. Here, $\mathcal{D}_k(Z_t)$ denotes the distribution of the chain (Z_t) at time t when started at $Z_0 = (k, 0)$ and $\pi_{Z_x} := \pi_{Z_x}^{-1}$ is the stationary distribution of this chain.

Proof: Fix $X_0 = x$ with $S(x) = k$ and consider the equivalence relation \sim_x corresponding to the classes $L(l) \cap D(x, l')$, for $l, l' \in \{0, 1, \dots, n\}$. That is, for $y, z \in \{0, 1\}^n$ we have

$$y \sim_x z \Leftrightarrow S(y) = S(z) \text{ and } d(x, y) = d(x, z).$$

Then the projection (Z_t) of (X_t) is a Markov chain if

$$P(y, L(h) \cap D(x, h')) = P(z, L(h) \cap D(x, h')) \quad (16)$$

for all $h, h' \in \{0, 1, \dots, n\}$ and all $y, z \in \{0, 1\}^n$ such that $y \sim_x z$. But this follows from symmetry (exchangeability of the coordinates) in our Markov chain (X_t) . To see this, fix any $y, z \in \{0, 1\}^n$

such that $y \sim_x z$ and let $l := S(y) = S(z)$ and $l' := d(x, y) = d(x, z)$. Fix any $h, h' \in \{0, 1, \dots, n\}$. Note that $S(X_t)$ and $d(x, X_t)$ can only stay the same or change by ± 1 in one step of the Markov chain. So in order to show (16), there are only a few cases to distinguish:

$l = h$: Equation (16) reads “ $P(y, y) = P(z, z)$ ” for $l' = h'$ and “ $0 = 0$ ” for $l' \neq h'$.

$l = h - 1$: For $l' = h' - 1$ equation (16) reads

$$P\{\text{pick } i \in G(x, y) \text{ and flip } 0 \rightarrow 1\} = \#G(x, y) \frac{1}{2n} \theta = \#G(x, z) \frac{1}{2n} \theta = P\{\text{pick } i \in G(x, z) \text{ and flip } 0 \rightarrow 1\}.$$

Similarly for $l' = h' + 1$. For $l' \neq h' \pm 1$ we get “ $0 = 0$ ”.

$l = h + 1$: This is entirely analogous to the case $l = h - 1$.

In all other cases equation (16) reads “ $0 = 0$ ”. This proves (16), showing that the process (Z_t) is in fact a Markov chain.

Total variation distance to stationarity remains unchanged under the projection Z_x because both $P^t(x, \cdot)$ and π are constant on sets $L(l) \cap D(x, l')$ for $l, l' \in \{0, 1, \dots, n\}$ by Proposition 5.1. That allows us to pull out the absolute values from the inner sum in the second equation below, because all the terms in the sum are equal:

$$\begin{aligned} \|P^t(x, \cdot) - \pi\| &= \frac{1}{2} \sum_l \sum_{l'} \sum_{z \in L(l) \cap D(x, l')} |P^t(x, z) - \pi(z)| \\ &= \frac{1}{2} \sum_l \sum_{l'} \left| \sum_{z \in L(l) \cap D(x, l')} P^t(x, z) - \pi(z) \right| \\ &= \frac{1}{2} \sum_l \sum_{l'} |P^t(x, \cdot) Z_x^{-1}(l, l') - \pi Z_x^{-1}(l, l')| \\ &= \|P^t(x, \cdot) Z_x^{-1} - \pi Z_x^{-1}\|. \end{aligned}$$

Clearly, $P^t(x, \cdot) Z_x^{-1} = \mathcal{D}_k(Z_t)$ and the fact that $\pi_{Z_x} := \pi Z_x^{-1}$ is stationary for (Z_t) is an elementary calculation summarized in the Lemma below. \square

Lemma 5.3. *Let (X_t) be a Markov chain on \mathcal{X} and suppose $S : \mathcal{X} \rightarrow \mathcal{Y}$ is onto and such that*

$$P(x, S^{-1}(y)) = P(x', S^{-1}(y))$$

for all $y \in \mathcal{Y}$ and all $x, x' \in \mathcal{X}$ such that $S(x) = S(x')$, so that $(S(X_t))$ is a Markov chain on \mathcal{Y} . If π is stationary for (X_t) , then πS^{-1} is stationary for $(S(X_t))$.

Proof: Fix $y \in \mathcal{Y}$ and for every $z \in \mathcal{Y}$ pick some $v_z \in S^{-1}(z)$. Then by stationarity of π for (X_t) we get

$$\sum_{z \in \mathcal{Y}} \pi S^{-1}(z) P(z, y) = \sum_{z \in \mathcal{Y}} \pi S^{-1}(z) P(v_z, S^{-1}(y))$$

$$\begin{aligned}
&= \sum_{z \in \mathcal{Y}} \sum_{x \in S^{-1}(z)} \pi(x) P(v_z, S^{-1}(y)) \\
&= \sum_{z \in \mathcal{Y}} \sum_{x \in S^{-1}(z)} \pi(x) P(x, S^{-1}(y)) \\
&= \sum_{x \in \mathcal{X}} \pi(x) P(x, S^{-1}(y)) \\
&= \sum_{x \in \mathcal{X}} \pi(x) \sum_{v \in S^{-1}(y)} P(x, v) \\
&= \sum_{v \in S^{-1}(y)} \sum_{x \in \mathcal{X}} \pi(x) P(x, v) \\
&= \sum_{v \in S^{-1}(y)} \pi(v) \\
&= \pi S^{-1}(y). \quad \square
\end{aligned}$$

Remark 5.4: Proposition 5.1, the Corollary 5.2 and their proofs stay true when we allow $\theta = \theta_{S(X_t)}$ to depend on $S(X_t)$, as in the random walk Metropolis chain for an arbitrary rotationally symmetric (i.e. $\pi(x) = \pi(y)$ whenever $S(x) = S(y)$) distribution π on the hypercube, as discussed in the introduction. Also note that unimodality of π is not needed for this result.

To establish the upper bound part of cutoff in Theorem 2.1, by the Corollary it remains to show that for $t := t_{n,\alpha} := \frac{1}{1+\theta} n \log n + \alpha n$ we get

$$\lim_{\alpha \rightarrow \infty} \limsup_{n \rightarrow \infty} \sup_{x \in \{0,1\}^n} \|\mathcal{D}_{(S(x))} Z_t - \pi_{Z_x}\| = 0. \quad (17)$$

For any fixed $X_0 = x$ with $S(x) = k$ the projected chain $(Z_t) = (Z_x(X_t))$ has the following transition kernel:

$$p((l, l'), (h, h')) = \begin{cases} \frac{2n-(l'+l+k)}{4n} \theta & : h = l+1, h' = l'+1, \\ \frac{l'+l-k}{4n} & : h = l-1, h' = l'-1, \\ \frac{k+l'-l}{4n} \theta & : h = l+1, h' = l'-1, \\ \frac{k-(l'-l)}{4n} & : h = l-1, h' = l'+1, \\ \frac{1}{2} + \frac{n-l}{2n} (1-\theta) & : h = l, h' = l', \\ 0 & : \text{otherwise.} \end{cases} \quad (18)$$

This follows from (15) together with the transition rule of the original chain (X_t) . For example,

$$p((l, l'), (l+1, l'+1)) = P\{\text{pick } i \in G \text{ and flip } 0 \rightarrow 1\} = \#G \frac{1}{2n} \theta.$$

For $k \leq \frac{n}{2}$ the state space of this chain is

$$\begin{aligned} \{(l, l') \in \{0, 1, \dots, n\}^2 : & \quad l' \in k + \{-l, -l+2, \dots, l\} & : & \quad \text{for } l \leq k, \\ & \quad l' \in k + \{l-2k, l-2k+2, \dots, l\} & : & \quad \text{for } k < l < n-k, \\ & \quad l' \in n-k + \{l-n, l-n+2, \dots, n-l\} & : & \quad \text{for } n-k \leq l\}. \end{aligned}$$

A similar result holds for $k \geq \frac{n}{2}$. In both cases, after reparametrizing

$$(l, l') \mapsto (l' - l, l' + l) =: (r, r'),$$

the state space becomes $\{(r, r') : r \in \{-k, -k+2, \dots, k\}, r' \in \{k, k+2, \dots, 2n-k\}\}$. The boundaries $-k \leq r \leq k$ and $k \leq r' \leq 2n-k$ here can also be confirmed like this: By definition we have

$$\begin{aligned} r &= l' - l = N + E - (N + F) = E - F & \text{and} \\ r' &= l' + l = N + E + (N + F) = 2N + k. \end{aligned}$$

Since $F \geq 0$, we get $r = E - F \leq E \leq E + F = k$. Also, since $E \geq 0$ and $F \leq k$, we get $r = E - F \geq -F \geq -k$. Similarly, since $N \leq n - k$, we get $r' = 2N + k \leq 2(n - k) + k = 2n - k$. And since $N \geq 0$, we get $r' = 2N + k \geq k$.

The transition kernel in this new parametrization becomes

$$p((r, r'), (s, s')) = \begin{cases} \frac{2n-(r'+k)}{4n}\theta & : \quad s = r, \quad s' = r' + 2, \\ \frac{r'-k}{4n} & : \quad s = r, \quad s' = r' - 2, \\ \frac{k+r}{4n}\theta & : \quad s = r - 2, \quad s' = r', \\ \frac{k-r}{4n} & : \quad s = r + 2, \quad s' = r', \\ \frac{1}{2} + \frac{2n-(r'-r)}{4n}(1-\theta) & : \quad s = r, \quad s' = r', \\ 0 & : \quad \text{otherwise.} \end{cases} \quad (19)$$

So the chain (Z_t) can be viewed as a birth and death chain on a rectangle in \mathbb{Z}^2 . A particular feature of this chain is that the probability of moving up (down) in the r -dimension only depends on the current location in that dimension: it only depends on r , not on r' . Similarly, the probability of moving up (down) in the r' -dimension only depends on r' , not on r . The problem of coupling two versions of this chain can therefore be split up into coupling two one-dimensional chains. Note also that this feature will be lost if we allow $\theta = \theta_{S(X_t)}$ to depend on the number of ones in the current state X_t .

We begin by calculating the expected location of the chain (Z_t) after t steps when started at $Z_0 = (r, r')$. Similar to the one-dimensional projection $S(X_t)$ that we analyzed for the lower bound, this can be calculated explicitly by induction on t since the transition probabilities (19) are all linear in the current location (r, r') of the chain.

For this, fix $X_0 = x \in \{0, 1\}^n$ and let $k = S(x)$, $Z_t = Z_x(X_t) = (S(X_t), d(x, X_t))$, so that $Z_0 = (k, 0)$. Denote with $Z_t = (Z_t^{(r)}, Z_t^{(r')})$ the coordinates of the chain in the new parametrization (19), so that $(Z_0^{(r)}, Z_0^{(r')}) = (-k, k)$, and write E_k for the expectation operator given this starting state. Then for this parametrization,

$$Z_{t+1} - Z_t = \begin{cases} (0, 2) & \text{with probability } \frac{2n - (k + Z_t^{(r')})}{4n} \theta, \\ (0, -2) & \text{with probability } \frac{Z_t^{(r')} - k}{4n}, \\ (-2, 0) & \text{with probability } \frac{k + Z_t^{(r)}}{4n} \theta, \\ (2, 0) & \text{with probability } \frac{k - Z_t^{(r)}}{4n}, \\ (0, 0) & \text{otherwise.} \end{cases}$$

Therefore,

$$\begin{aligned} E[Z_{t+1} - Z_t \mid Z_t] &= \left(-2 \frac{k + Z_t^{(r)}}{4n} \theta + 2 \frac{k - Z_t^{(r)}}{4n}, 2 \frac{2n - (k + Z_t^{(r')})}{4n} \theta - 2 \frac{Z_t^{(r')} - k}{4n} \right) \\ &= \left(\frac{k(1 - \theta) - Z_t^{(r)}(1 + \theta)}{2n}, \frac{2n\theta + k(1 - \theta) - Z_t^{(r')}(1 + \theta)}{2n} \right), \end{aligned}$$

so that

$$E[Z_{t+1} \mid Z_t] = \left(\frac{k(1 - \theta) + [2n - (1 + \theta)]Z_t^{(r)}}{2n}, \frac{2n\theta + k(1 - \theta) + [2n - (1 + \theta)]Z_t^{(r')}}{2n} \right).$$

By taking expectation, we get

$$\begin{aligned} E_k[Z_{t+1}] &= \left(\frac{k(1 - \theta) + [2n - (1 + \theta)]E_k Z_t^{(r)}}{2n}, \frac{2n\theta + k(1 - \theta) + [2n - (1 + \theta)]E_k Z_t^{(r')}}{2n} \right) \\ &=: \left(\beta + \gamma E_k Z_t^{(r)}, \theta + \beta + \gamma E_k Z_t^{(r')} \right). \end{aligned} \tag{20}$$

By induction on t , this leads to a proof of the following result:

Proposition 5.5. *Let $Z_t = Z_x(X_t) = (S(X_t), d(x, X_t))$ be the two-dimensional projection of the lazy random walk Metropolis chain (X_t) for $\pi(x) = \theta^{S(x)}(1 + \theta)^{-n}$, started at $X_0 = x \in \{0, 1\}^n$ with $S(x) = k$. Then in the parametrization (19) and for any $t \in \mathbb{N}$ we get*

$$E_k \left(Z_t^{(r)}, Z_t^{(r')} \right) = \left(\frac{2n\beta}{1 + \theta} (1 - \gamma^t) - k\gamma^t, \frac{2n(\theta + \beta)}{1 + \theta} (1 - \gamma^t) + k\gamma^t \right),$$

where $\beta := \beta_{n,k,\theta} := \frac{k}{2n}(1 - \theta)$ and $\gamma := \gamma_{n,\theta} := 1 - \frac{1+\theta}{2n}$.

Proof: The claim is true for $t = 1$, since by (20) we get

$$\begin{aligned} E_k \left(Z_1^{(r)}, Z_1^{(r')} \right) &= (\beta - \gamma k, \theta + \beta + \gamma k) \\ &= \left(\frac{2n\beta}{1+\theta} \frac{1+\theta}{2n} - \gamma k, \frac{2n(\theta+\beta)}{1+\theta} \frac{1+\theta}{2n} + \gamma k \right). \end{aligned}$$

Now suppose the claim is true for t . Then by (20) we get that $E_k \left(Z_{t+1}^{(r)}, Z_{t+1}^{(r')} \right)$ is equal to

$$\begin{aligned} &\left(\beta + \gamma \left[\frac{2n\beta}{1+\theta} (1 - \gamma^t) - k\gamma^t \right], \theta + \beta + \gamma \left[\frac{2n(\theta+\beta)}{1+\theta} (1 - \gamma^t) + k\gamma^t \right] \right) \\ &= \left(\beta + \frac{2n\beta}{1+\theta} \gamma - \frac{2n\beta}{1+\theta} \gamma^{t+1} - k\gamma^{t+1}, \theta + \beta + \frac{2n(\theta+\beta)}{1+\theta} \gamma - \frac{2n(\theta+\beta)}{1+\theta} \gamma^{t+1} + k\gamma^{t+1} \right) \\ &= \left(\frac{2n\beta}{1+\theta} \left[\frac{1+\theta}{2n} + \gamma \right] - \frac{2n\beta}{1+\theta} \gamma^{t+1} - k\gamma^{t+1}, \frac{2n(\theta+\beta)}{1+\theta} \left[\frac{1+\theta}{2n} + \gamma \right] - \frac{2n(\theta+\beta)}{1+\theta} \gamma^{t+1} + k\gamma^{t+1} \right) \\ &= \left(\frac{2n\beta}{1+\theta} (1 - \gamma^{t+1}) - k\gamma^{t+1}, \frac{2n(\theta+\beta)}{1+\theta} (1 - \gamma^{t+1}) + k\gamma^{t+1} \right). \end{aligned}$$

So it is also true for $t + 1$, finishing the proof. \square

Corollary 5.6. *For the expectation under stationarity in the parametrization (19) of our two-dimensional chain (Z_t) we get*

$$E_\pi \left(Z^{(r)}, Z^{(r')} \right) = \left(\frac{2n\beta}{1+\theta}, \frac{2n(\theta+\beta)}{1+\theta} \right) = \left(k \frac{1-\theta}{1+\theta}, \frac{2n\theta}{1+\theta} + k \frac{1-\theta}{1+\theta} \right).$$

Proof: Since the Markov chain (Z_t) is irreducible and aperiodic, it converges to its unique stationary distribution as t goes to infinity for fixed n, k . Since the state space is finite, this convergence also holds in L_1 . So by the Proposition, we get

$$\begin{aligned} E_\pi \left(Z^{(r)}, Z^{(r')} \right) &= \lim_{t \rightarrow \infty} E_k \left(Z_t^{(r)}, Z_t^{(r')} \right) \\ &= \left(\frac{2n\beta}{1+\theta}, \frac{2n(\theta+\beta)}{1+\theta} \right), \end{aligned}$$

since for fixed n, k we have $\gamma^t \rightarrow 0$ as t goes to infinity. \square

Remark 5.7: By reversing the linear transformation $(l, l') \mapsto (l' - l, l' + l)$, we immediately get

$$E_\pi Z = \left(\frac{n\theta}{1+\theta}, \frac{n\theta + k(1-\theta)}{1+\theta} \right)$$

for the expectation under stationarity in the original parametrization (18) of our two-dimensional chain (Z_t) . For the first coordinate this confirms what we already know from $S(X) \sim \text{Binomial}(n, \frac{\theta}{1+\theta})$ under stationarity.

Remark 5.8: By exactly the same proof we get for a general starting state (r, r') in the new parametrization (19) that

$$E_{(r, r')} \left(Z_t^{(r)}, Z_t^{(r')} \right) = \left(\frac{2n\beta}{1+\theta}(1 - \gamma^t) + r\gamma^t, \frac{2n(\theta + \beta)}{1+\theta}(1 - \gamma^t) + r'\gamma^t \right).$$

Now from Remark 3.3 on page 7 we know that for $u := u_{n, \theta} := \frac{1}{1+\theta}n \log n$ we get $\gamma^u \sim n^{-1/2}$, meaning that the ratio of these two quantities goes to one as n goes to infinity. Therefore, Proposition 5.5 and its Corollary imply that for any starting state the expected location of the chain (Z_t) after u steps is within $O(\sqrt{n})$ of the expected location of the chain under stationarity. To see this, just subtract the stationary expectation from the expectation after u steps and note that both $\frac{2n\beta}{1+\theta} + k$ and $\frac{2n(\theta + \beta)}{1+\theta} - k$ are in $\Theta(n)$. We now want to show that an additional αn number of steps is enough to couple two chains that are at distance $O(\sqrt{n})$ of their stationary mean. This will follow from a corresponding result for lazy simple random walk on \mathbb{Z}^2 , since close to the stationary mean we are now in the “entropy regime” where the drift of the chain is negligible so it behaves similar to lazy simple random walk.

For a pair of lazy simple random walks on \mathbb{Z}^2 we can couple the two coordinates of the chains “one by one” in a coupling of Ornstein type as follows: Let $V_t = (V_t^{(r)}, V_t^{(r')})$ and $W_t = (W_t^{(r)}, W_t^{(r')})$ be two versions of lazy simple random walk on \mathbb{Z}^2 , i.e. the transition probabilities are $p(x, x) = 1/2$ and $p(x, y) = 1/8$ for each of the four neighbors y of x . Without loss of generality suppose that $V_0^{(r)} \neq W_0^{(r)}$ and $V_0^{(r')} \neq W_0^{(r')}$.

Phase I: $(V_t^{(r)} \neq W_t^{(r)})$ We try to couple the r -coordinates by running an independence coupling for this coordinate while moving the r' -coordinates in “lockstep”, leaving $|V_t^{(r')} - W_t^{(r')}|$ constant: Flip fair coin number one. If it comes up heads, we move in the r -coordinate: Flip another fair coin to decide which of the two chains to move according to one-dimensional (non-lazy) simple random walk for the r -coordinate. The other chain stays at its current location. If fair coin number one comes up tails, we move in the r' -coordinate: Flip another fair coin. If it comes up heads, both chains stay where they are; if it comes up tails, both chains move in the *same* direction on the r' -coordinate according to (non-lazy) simple random walk. Note that $|V_t^{(r)} - W_t^{(r)}|$ performs a one-dimensional lazy simple random walk, while $|V_t^{(r')} - W_t^{(r')}|$ stays constant. Run in Phase I until $V_t^{(r)} = W_t^{(r)}$.

Phase II: $(V_t^{(r)} = W_t^{(r)})$ Now we switch the role of the two coordinates: We try to couple the r' -coordinates by running an independence coupling for this coordinate, while moving the r -coordinates in “lockstep”, leaving $|V_t^{(r)} - W_t^{(r)}| = 0$. Flip fair coin number one. If it comes up heads, we move in the r' -coordinate: Flip another fair coin to decide which of the two chains to move according to one-dimensional (non-lazy) simple random walk for the r' -coordinate. The other chain stays at its current location. If fair coin number one comes up tails, we move in the

r -coordinate: Flip another fair coin. If it comes up heads, both chains stay where they are; if it comes up tails, both chains move in the *same* direction on the r -coordinate according to (non-lazy) simple random walk. Note that $|V_t^{(r')} - W_t^{(r')}|$ performs a one-dimensional lazy simple random walk, while $|V_t^{(r)} - W_t^{(r)}|$ stays constant equal to zero. Run in this phase till $V_t^{(r')} = W_t^{(r')}$ and therefore $V_t = W_t$. \square

So coupling a pair of two-dimensional (lazy) simple random walks can be achieved by coupling two pairs of one-dimensional (lazy) simple random walks, one after the other. Therefore we start with the following result for one-dimensional chains:

Proposition 5.9. *Let (V_t) be $(1-2\delta)$ -lazy simple random walk on \mathbb{Z} , started at $V_0 = k$. That is, for $\delta \in (0, \frac{1}{2}]$ we have $V_t = k + \sum_{i=1}^t \xi_i$ where the ξ_i are iid with $P\{\xi_i = \pm 1\} = \delta$, $P\{\xi_i = 0\} = 1 - 2\delta$. Let $\tau_0 := \min\{t' \geq 0 : V_{t'} = 0\}$ be the first time the walk hits zero. Then there exists a constant C such that for all $k \in \mathbb{Z}$ and $r \in \mathbb{N}$ we get*

$$P_k\{\tau_0 > r\} \leq \frac{C|k|}{\sqrt{r\delta}}. \quad (21)$$

This is a straightforward generalization of the corresponding well known result for simple random walk ($\delta = \frac{1}{2}$). For completeness, we give a proof below. We follow the proof of Corollary 2.28 on page 36 in Levin et al. (2009) for the case where $\delta = \frac{1}{4}$. It is a consequence of the following result (see Theorem 2.26 on page 35 in Levin et al. (2009) for a proof).

Theorem 5.10. *Let (ξ_i) be iid integer-valued random variables with mean zero and variance σ^2 . Let $X_t = \sum_{i=1}^t \xi_i$, with $X_0 = 0$. Then*

$$P\{X_t \neq 0 \text{ for } 1 \leq t \leq r\} \leq \frac{4\sigma}{\sqrt{r}}.$$

Proof of Proposition 5.9: For $k = 0$ the claim is true, so by symmetry we may assume that $k \geq 1$. Define $\tau_0^+ := \min\{t' \geq 1 : V_{t'} = 0\}$ and $\tau_k := \min\{t' \geq 0 : V_{t'} = k\}$ for $k \in \mathbb{Z}$. Then by conditioning on the first step of the walk started at zero, we get from symmetry that

$$\begin{aligned} P_0\{\tau_0^+ > r\} &= \delta P_1\{\tau_0^+ > r-1\} + \delta P_{-1}\{\tau_0^+ > r-1\} \\ &= 2\delta P_1\{\tau_0^+ > r-1\}. \end{aligned}$$

Therefore

$$\begin{aligned} P_1\{\tau_k < \tau_0\} P_k\{\tau_0^+ > r\} &= P_1\{\tau_k < \tau_0 \text{ and don't hit zero for } r \text{ steps after hitting } k\} \\ &\leq P_1\{\tau_0^+ > r-1\} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2\delta} P_0\{\tau_0^+ > r\} \\
&\leq \frac{1}{2\delta} \frac{4\sqrt{\text{Var}(\xi_i)}}{\sqrt{r}} \\
&= \frac{2\sqrt{2}}{\sqrt{r\delta}}.
\end{aligned}$$

Here, the first inequality follows since the first event is contained in the second. The second inequality follows from Theorem 5.10. Since $k \geq 1$, we get $P_1\{\tau_k < \tau_0\} = \frac{1}{k}$ by Gamblers ruin, so

$$\begin{aligned}
P_k\{\tau_0 > r\} &= P_k\{\tau_0^+ > r\} \\
&\leq \frac{2\sqrt{2}k}{\sqrt{r\delta}}. \quad \square
\end{aligned}$$

Next we show that by running the chain for an extra αn steps (burn-in period), we may assume that the number of ones in the state that is used in the two-dimensional projection is close to its stationary mean. Fix $\delta > 0$ and let $p := \frac{\theta}{1+\theta}$. We will show that

$$\max_{x \in \{0,1\}^n} \|P^{t+\alpha n}(x, \cdot) - \pi\| \leq \max_{y: S(y) \in n(p \pm \delta)} \|P^t(y, \cdot) - \pi\| + o(1), \quad (22)$$

where the $o(1)$ term goes to zero as n goes to infinity. To see this, we condition on where we are after the first αn steps:

$$\begin{aligned}
\|P^{t+\alpha n}(x, \cdot) - \pi\| &= \left\| \sum_y P^{\alpha n}(x, y) [P^t(y, \cdot) - \pi] \right\| \\
&\leq \sum_{y: S(y) \in n(p \pm \delta)} P^{\alpha n}(x, y) \|P^t(y, \cdot) - \pi\| + \sum_{y: S(y) \notin n(p \pm \delta)} P^{\alpha n}(x, y) \|P^t(y, \cdot) - \pi\| \\
&\leq \max_{y: S(y) \in n(p \pm \delta)} \|P^t(y, \cdot) - \pi\| + P_x\{S(X_{\alpha n}) \notin n(p \pm \delta)\}.
\end{aligned}$$

The last term on the right hand side is in fact in $o(1)$ because when writing $S_{\alpha n} := S(X_{\alpha n})$ we get for large α that

$$\begin{aligned}
P_x\{S(X_{\alpha n}) \notin n(p \pm \delta)\} &= P_x\{|S_{\alpha n} - E_x S_{\alpha n} + E_x S_{\alpha n} - np| \geq \delta n\} \\
&\leq P_x\{|S_{\alpha n} - E_x S_{\alpha n}| \geq \delta n - |E_x S_{\alpha n} - np|\} \\
&\leq P_x\left\{|S_{\alpha n} - E_x S_{\alpha n}| \geq n \left(\delta - \exp\left\{-\frac{1+\theta}{2}\alpha\right\}\right)\right\} \\
&\leq 2 \exp\left\{-\frac{2n^2 \left(\delta - \exp\left\{-\frac{1+\theta}{2}\alpha\right\}\right)^2}{9\alpha n}\right\} \\
&= o(1).
\end{aligned}$$

Here the second inequality holds since we get from Proposition 3.2 that $|E_x S_{\alpha n} - np| \leq n\gamma^{\alpha n} \leq n \exp\{-\frac{1+\theta}{2}\alpha\}$. The third inequality follows from our concentration result based on Azuma's inequality (Proposition 4.4). This proves (22). So after an initial αn steps we may assume that the number of ones is within δn of its stationary mean np .

Now we project down to our two-dimensional chain: Fix any $y \in \{0, 1\}^n$ such that $k := S(y) \in n(p \pm \delta)$. Writing the transition kernel in the parametrization (19) we get from Corollary 5.2 that

$$\begin{aligned}
\|P^t(y, \cdot) - \pi\| &= \|\mathcal{D}_{(S(y))} Z_t - \pi_{Z_y}\| \\
&= \|P^t((-k, k), \cdot) - \pi_{Z_y}\| \\
&\leq \max_{(v, v')} \|P^t((v, v'), \cdot) - \pi_{Z_y}\| \\
&\leq \max_{(v, v'), (w, w')} \|P^t((v, v'), \cdot) - P^t((w, w'), \cdot)\| \\
&\leq \max_{(v, v'), (w, w')} P_{(v, v'), (w, w')} \{\tau > t\}.
\end{aligned} \tag{23}$$

The maxima here are over the entire state space of the two-dimensional chain $(v, v'), (w, w') \in \{-k, \dots, k\} \times \{k, \dots, 2n - k\}$. The second inequality above is well known, see for example Lemma 4.11 in Levin et al. (2009, page 53). The last inequality is the coupling inequality, where $\tau := \min\{j \geq 0 : Z_j = Y_j\}$ is the coupling time in the coupling (Z_j, Y_j) that we now describe.

Fix any $(v, v'), (w, w') \in \{-k, \dots, k\} \times \{k, \dots, 2n - k\}$ and set $Z_0 := (v, v')$ and $Y_0 := (w, w')$. Let $t = s + u$, where $s := \frac{1}{1+\theta}n \log n$ and $u := \alpha n$. For steps $j = 1, 2, \dots, s$ we use an independence coupling, i.e. at each step we flip a fair coin to decide which chain to move according to the non-lazy version of its transition kernel. The other chain stays at its current location. Here, if P is the transition probability matrix (19) of the chain (Z_j) , then $P' := 2P - I$ is its non-lazy version, where I is the identity. However, if Z_j and Y_j ever agree in the r (or r') coordinate, we modify the coupling so that they agree in that coordinate forever after. This is possible since the probability of moving up (or down) in the r -coordinate does not depend on the current location in the r' -coordinate. Similarly, the probability of moving up (or down) in the r' -coordinate does not depend on the current location in the r -coordinate. This is easily seen from the transition kernel (19).

We could implement this change as follows: Suppose $Y_j^{(r)} = Z_j^{(r)}$. Flip fair coin number one; if it comes up heads, try to move Y_j according to its non-lazy transition rule. If that would result in Y_j moving up (or down) in the r -coordinate, flip another fair coin. If it comes up heads, move $Y_j^{(r)}$ accordingly and move $Z_j^{(r)}$ in the same way; if it comes up tails, reject the move. If fair coin number one comes up tails, try to move Z_j according to its non-lazy transition rule. If that would result in Z_j moving up (or down) in the r -coordinate, flip another fair coin. If it comes up heads, move $Z_j^{(r)}$ accordingly and move $Y_j^{(r)}$ in the same way; if it comes up tails, reject the move. Similarly for

$$Y_j^{(r')} = Z_j^{(r')}.$$

For steps $j = s + 1, s + 2, \dots, s + u$ we couple (Z_j, Y_j) with an Ornstein-type coupling (V_j, W_j) of two lazy simple random walks on \mathbb{Z}^2 as described above, where now the laziness-factor (the probability of not moving) at step j depends on the current location of (Z_j, Y_j) . Write $D_j^{(r)} := Z_j^{(r)} - Y_j^{(r)}$ and $D_j^{(r')} := Z_j^{(r')} - Y_j^{(r')}$ for the distance between Z_j and Y_j in the r and r' coordinates at time j , respectively. Similarly we write $C_j^{(r)} := V_j^{(r)} - W_j^{(r)}$ and $C_j^{(r')} := V_j^{(r')} - W_j^{(r')}$. To start, let $V_s := Z_s$ and $W_s := Y_s$, so that $D_s^{(r)} = C_s^{(r)}$ and $D_s^{(r')} = C_s^{(r')}$.

Note that the chains $(Z_j), (Y_j)$ have a (small) drift towards their stationary mean. A short calculation based on (19) and Corollary 5.6 shows that this drift is linear in the distance to the stationary mean. Since $(V_j), (W_j)$ don't have a drift, this coupling can be done in such a way that at all times and in both coordinates, the distance between Z_j and Y_j is no greater than the distance between V_j and W_j . Throughout, once the two chains meet (in the r or r' coordinate), we let them stay together in that coordinate forever, i.e. if $Y_j^{(r)} = Z_j^{(r)}$ then $Y_i^{(r)} = Z_i^{(r)}$ for all $i \geq j$. Similarly for the r' -coordinate. This can be implemented as follows:

Phase I: ($V_j^{(r)} \neq W_j^{(r)}$) We first try to couple the r -coordinates of V_j and W_j by running an independence coupling for this coordinate while moving the r' -coordinates in “lockstep”, leaving $|C_j^{(r')}| = |V_j^{(r')} - W_j^{(r')}|$ constant. To this end, let

$$\rho_j := \frac{P\{Y_j^{(r)} \text{ moves} \mid Y_j\} + P\{Z_j^{(r)} \text{ moves} \mid Z_j\}}{2} = \frac{k}{n} \frac{1 + \theta}{4} - \frac{Y_j^{(r)} + Z_j^{(r)}}{2n} \frac{1 - \theta}{4} \quad (24)$$

be the average of the probabilities that Y_j respectively Z_j moves in the r -coordinate. Note that ρ_j is random and depends on the current location of $Y_j^{(r)}$ and $Z_j^{(r)}$. Throw a $\text{Ber}(2\rho_j)$ coin. If it comes up heads, we move in the r -coordinate:

- If $Y_j^{(r)} \neq Z_j^{(r)}$, use one $\text{Uniform}[0, 1]$ random variable to pick one of the four possible moves $Y_j^{(r)}$ up, $Y_j^{(r)}$ down, $Z_j^{(r)}$ up, $Z_j^{(r)}$ down; let the same $\text{Uniform}[0, 1]$ variable determine one of the four possible moves $V_j^{(r)}$ up, $V_j^{(r)}$ down, $W_j^{(r)}$ up, $W_j^{(r)}$ down. This can be done in such a way that $|D_{j+1}^{(r)}| \leq |C_{j+1}^{(r)}|$.
- If $Y_j^{(r)} = Z_j^{(r)}$ we have $\rho_j = P\{Y_j^{(r)} \text{ moves} \mid Y_j\} = P\{Z_j^{(r)} \text{ moves} \mid Z_j\}$ and we can move $Y_j^{(r)}$ and $Z_j^{(r)}$ either both up, or both down, or both stay. At the same time, independently pick one of the four possible moves $V_j^{(r)}$ up, $V_j^{(r)}$ down, $W_j^{(r)}$ up, $W_j^{(r)}$ down with equal probability.

If our $\text{Ber}(2\rho_j)$ coin comes up tails, we move in the r' -coordinate (or we don't move). This can be done in such a way that $|C_{j+1}^{(r')}| = |C_j^{(r')}|$ and $|D_{j+1}^{(r')}| \leq |D_j^{(r')}|$. Overall, this coupling in Phase I ensures that $|D_j| \leq |C_j|$ for all $j = 1, 2, \dots$ and in both coordinates r and r' . Furthermore, $(|C_j^{(r)}|)$

performs a lazy simple random walk where the probability of not moving is $1 - 2\rho_j$; this probability is a random variable that depends on the path of $(Y_j^{(r)}, Z_j^{(r)})$.

Phase II: ($V_j^{(r)} = W_j^{(r)}$) We have the r -coordinates matched, $D_j^{(r)} = C_j^{(r)} = 0$, and want to keep them that way. At the same time, we try to couple the r' -coordinates. To this end, let

$$\sigma_j := \frac{P\{Y_j^{(r')} \text{ moves} \mid Y_j\} + P\{Z_j^{(r')} \text{ moves} \mid Z_j\}}{2} = \frac{\theta}{2} - \frac{k}{n} \frac{1+\theta}{4} + \frac{Z_j^{(r')} + Y_j^{(r')}}{2n} \frac{1-\theta}{4} \quad (25)$$

be the average of the probabilities that Y_j respectively Z_j moves in the r' -coordinate. Note that σ_j is random and depends on the current location of $Y_j^{(r')}$ and $Z_j^{(r')}$. Throw a $\text{Ber}(2\sigma_j)$ coin. If it comes up heads, we move in the r' -coordinate:

- If $Y_j^{(r')} \neq Z_j^{(r')}$, use one $\text{Uniform}[0, 1]$ random variable to pick one of the four possible moves $Y_j^{(r')}$ up, $Y_j^{(r')}$ down, $Z_j^{(r')}$ up, $Z_j^{(r')}$ down; let the same $\text{Uniform}[0, 1]$ variable determine one of the four possible moves $V_j^{(r')}$ up, $V_j^{(r')}$ down, $W_j^{(r')}$ up, $W_j^{(r')}$ down. This can be done in such a way that $|D_{j+1}^{(r')}| \leq |C_{j+1}^{(r')}|$.
- If $Y_j^{(r')} = Z_j^{(r')}$ we have $\sigma_j = P\{Y_j^{(r')} \text{ moves}\} = P\{Z_j^{(r')} \text{ moves}\}$ and we can move $Y_j^{(r')}$ and $Z_j^{(r')}$ either both up, or both down, or both stay. At the same time, independently pick one of the four possible moves $V_j^{(r')}$ up, $V_j^{(r')}$ down, $W_j^{(r')}$ up, $W_j^{(r')}$ down with equal probability.

If our $\text{Ber}(2\sigma_j)$ coin comes up tails, we move in the r -coordinate (or we don't move). This can be done in such a way that $|C_{j+1}^{(r)}| = |C_j^{(r)}| = 0$ and $|D_{j+1}^{(r)}| = |D_j^{(r)}| = 0$. Overall, this coupling in Phase II ensures that $|D_j^{(r')}| \leq |C_j^{(r')}|$ and $|D_j^{(r)}| = |C_j^{(r)}| = 0$ for all $j = 1, 2, \dots$. Furthermore, $(|C_j^{(r')}|)$ performs a lazy simple random walk, where the probability of not moving is the random variable $1 - 2\sigma_j$ that depends on the path of $(Y_j^{(r')}, Z_j^{(r')})$.

By (22) and (23), we may assume that $\frac{k}{n} \in \left(\frac{\theta}{1+\theta} \pm \delta\right)$ for any $\delta > 0$. As a consequence, the (random) laziness factors for the simple random walks $(|C_j^{(r)}|)$ and $(|C_j^{(r')}|)$ are bounded away from one, since in that case

$$\begin{aligned} P\{|C_j^{(r)}| \text{ moves} \mid Y_j, Z_j\} = 2\rho_j &\geq \frac{k}{n}\theta \geq \left(\frac{\theta}{1+\theta} - \delta\right)\theta =: 2\delta_r > 0, \\ P\{|C_j^{(r')}| \text{ moves} \mid Y_j, Z_j\} = 2\sigma_j &\geq \frac{n-k}{n}\theta \geq \left(\frac{1}{1+\theta} - \delta\right)\theta =: 2\delta_{r'} > 0, \end{aligned}$$

where we picked some $\delta \in (0, \frac{\theta}{1+\theta})$ to ensure the two strict inequalities. This follows from the definitions of ρ_j and σ_j by noting that the r -coordinates lie in $[-k, k]$ while the r' -coordinates lie in $[k, 2n - k]$.

Let $\tau := \min\{j \geq 0 : V_j = W_j\}$ be the coupling time of (V_j, W_j) . Denote with $\hat{\tau}_r$ and $\hat{\tau}_{r'}$ the time we spend in Phase I and Phase II of the coupling described above, so that $\tau = \hat{\tau}_r + \hat{\tau}_{r'}$. That means $\hat{\tau}_r$ (respectively $\hat{\tau}_{r'}$) is the time it takes for the r (respectively r') coordinates of V_j and W_j to couple, i.e. the time it takes for the lazy simple random walk $(|C_j^{(r)}|)$ (respectively $(|C_j^{(r')}|)$) to hit zero. (Again, the laziness factors here are random, depending on the path of (Z_j, Y_j) .)

Now we compare these two lazy one-dimensional simple random walks to two other one-dimensional simple random walks that are uniformly lazier, but with deterministic laziness factor. For the walk corresponding to the r -chain the probability of an upward move will be δ_r . For the walk corresponding to the r' -chain the probability of an upward move will be $\delta_{r'}$. Let $\tilde{\tau}_r$ (respectively $\tilde{\tau}_{r'}$) be the first time this deterministically lazy chain hits zero. By letting it follow the path of $(|C_j^{(r)}|)$ (respectively $(|C_j^{(r')}|)$) at a slower pace, we can ensure that $\hat{\tau}_r \leq \tilde{\tau}_r$ and $\hat{\tau}_{r'} \leq \tilde{\tau}_{r'}$ almost surely. The starting point of the walks will be indicated as a subscript in the probability measure.

Now we put things together. By conditioning on where we are after the first s steps (the drift regime of our coupling), we get

$$\begin{aligned}
P_{(v,v'),(w,w')}\{\tau > s + u \mid Z_s, Y_s\} &= \mathbf{1}\{\tau > s\} P_{Z_s, Y_s}\{\tau > u\} \\
&\leq P_{Z_s, Y_s}\{\hat{\tau} > u\} \\
&\leq P_{Z_s, Y_s}\{\hat{\tau}_r > u/2\} + P_{Z_s, Y_s}\{\hat{\tau}_{r'} > u/2\} \\
&\leq P_{D_s^{(r)}}\{\tilde{\tau}_r > u/2\} + P_{D_s^{(r')}}\{\tilde{\tau}_{r'} > u/2\} \\
&\leq \frac{C|D_s^{(r)}|}{\sqrt{\delta_r u}} + \frac{C|D_s^{(r')}|}{\sqrt{\delta_{r'} u}}.
\end{aligned}$$

Here the first inequality holds since $V_j = W_j$ implies $Z_j = Y_j$ by the construction of our coupling. The last inequality follows from Proposition 5.9 where C is a positive constant. By taking expectation, we get for large n that

$$\begin{aligned}
P_{(v,v'),(w,w')}\{\tau > s + u\} &\leq \frac{CE_{(v,v'),(w,w')}|D_s^{(r)}|}{\sqrt{\delta_r u}} + \frac{CE_{(v,v'),(w,w')}|D_s^{(r')}|}{\sqrt{\delta_{r'} u}} \\
&= \frac{C|E_{(v,v'),(w,w')}D_s^{(r)}|}{\sqrt{\delta_r u}} + \frac{C|E_{(v,v'),(w,w')}D_s^{(r')}|}{\sqrt{\delta_{r'} u}} \\
&= \frac{C|v - w|\gamma^s}{\sqrt{\delta_r u}} + \frac{C|v' - w'|\gamma^s}{\sqrt{\delta_{r'} u}} \\
&\leq \frac{C'n/\sqrt{n}}{\sqrt{\alpha n}} + \frac{C'n/\sqrt{n}}{\sqrt{\alpha n}} \\
&\leq \frac{\tilde{C}}{\sqrt{\alpha}}.
\end{aligned}$$

Here $E_{(v,v'),(w,w')}|D_s^{(r)}| = |E_{(v,v'),(w,w')}D_s^{(r)}|$ in the first equality follows since $D_0^{(r)} \geq 0$ implies $D_j^{(r)} \geq 0$ for all $j \leq s$ and $D_0^{(r)} \leq 0$ implies $D_j^{(r)} \leq 0$ for all $j \leq s$, by the construction of our coupling. Similarly for the r' -coordinate. (In both coordinates r and r' , the paths of Z_j and V_j never cross.) The second equality above uses Remark 5.8. In the second inequality we use the fact that $\gamma^s \sim 1/\sqrt{n}$, as shown in Remark 3.3.

Combining this with (22),(23), it follows for $t = s + u$ as above that for some constant \tilde{C} we get

$$\begin{aligned} \lim_{\alpha \rightarrow \infty} \limsup_{n \rightarrow \infty} \max_{x \in \{0,1\}^n} \|P^t(x, \cdot) - \pi\| &\leq \lim_{\alpha \rightarrow \infty} \limsup_{n \rightarrow \infty} \max_{(v,v'),(w,w')} P_{(v,v'),(w,w')} \{\tau > s + u\} \\ &\leq \lim_{\alpha \rightarrow \infty} \frac{\tilde{C}}{\sqrt{\alpha}} \\ &= 0. \end{aligned}$$

This finishes the proof of the upper bound part of Theorem 2.1 and we are done. \square

Acknowledgment: I would like to thank my advisor Steven Lalley for his invaluable guidance and many helpful discussions and suggestions.

6 References

- Diaconis, P. (1996). The Cutoff Phenomenon in Finite Markov Chains. *Proc Natl Acad Sci USA* 93(Feb), 1659-64.
- Diaconis, P., Graham R. L. and Morrison, J. A. (1990). Asymptotic Analysis of a Random Walk on a Hypercube with Many Dimensions. *Rand Struct Alg* 1(1), 51-72.
- Diaconis, P. and Hanlon, P. (1992). Eigen Analysis for Some Examples of the Metropolis Algorithm. *Contemporary Math* 138, 99-117.
- Diaconis, P. and Ram, A. (2000). Analysis of Systematic Scan Metropolis Algorithms Using Iwahori-Hecke Algebra Techniques. *Michigan Math J* 48, 157-90.
- Diaconis, P. and Saloff-Coste, L. (2006). Separation Cut-Offs for Birth and Death Chains. *Ann Appl Prob* 16(4), 2098-2122.
- Dubhashi, D. P. and Panconesi, A. (2009). *Concentration of Measure for the Analysis of Randomized Algorithms*. Cambridge: Cambridge University Press.
- Levin, D. A., Luczak, M. J. and Peres, Y. (2010). Glauber Dynamics for the Mean-Field Ising Model: Cutoff, Critical Power Law, and Metastability. *Prob Theory Rel Fields* 146, 223-65.
- Levin, D. A., Peres, Y. and Wilmer, E. L (2009). *Markov Chains and Mixing Times*. Providence, RI: AMS.
- Ross, S. (1996). *Stochastic Processes*, 2nd edition. New York: Wiley.
- Ross, K. A. and Xu, D. (1994). Hypergroup Deformations and Markov Chains. *J Theor Prob* 7(4), 813-30.